



**Hate
Aid**

**SA
FE
TY** **by DESIGN**

Pathways to Safer Social Media Platforms

Content

Shifting Paradigms: What is Safety by Design?	4
From Quick Fixes to Long-Term Solutions	4
The Principles of Safety by Design	7
Our Safety-by-Design Taxonomy	8
How to Implement Safety-by-Design Measures	11
Barriers to Implementation	11
Recommendations for Policymakers	13
List of Sources	16
Imprint	18

Supported by



Schöpflin Stiftung :

Shifting Paradigms: What is Safety by Design?

From Quick Fixes to Long-Term Solutions

In the European Union (EU), consumers rely on strong legal protections to ensure their safety in everyday life. Vehicle regulations require cars to be equipped with seat belts, airbags, and automatic emergency braking systems. Drug legislation demands that new medicines be thoroughly tested

before they are approved for sale. And child safety laws prohibit the use of certain chemicals in the fabrication of toys. At every stage of a product's life cycle—from design to manufacturing to distribution—consumer safety is the primary concern. This is known as “Safety by Design” (SbD).

Yet, when it comes to digital spaces like social media platforms, European laws are falling behind the rapid pace of technological change. Although these platforms have become essential hubs for public discourse, social interaction, and news consumption,¹ their design tends to prioritise engagement over user safety.² Algorithms peddle outrage, violence, and disinformation;³ autoplay and infinite scroll features lure children into social media addiction;⁴ and dopamine-reinforced rewards, such as “likes” and “reactions”, feed into anxiety and self-esteem issues.⁵

Though these adverse effects have been known for years, social media platforms remain reluctant to implement changes to their online appearances' design—fearing this will hurt their business model. Instead, they frequently tend to obfuscate evidence⁶ and arbitrarily apply reactive safety measures that can neither prevent online harms nor properly mitigate their consequences. While there is no doubt that measures like removing illegal content, factchecking misleading claims, or banning abusive users are crucial to user safety, their post factum approach has three key drawbacks:

- 1. Users must experience harm** on social media platforms before it will be noticed. Since most platforms do not properly assess potential harms, their users are likely to become living test subjects for the platforms' design decisions. A prominent example is X's recent changes to its chatbot Grok, which led to millions of cases of sexualised deepfakes.⁷ The lack of prevention is particularly dangerous for vulnerable users, such as children or people with trauma, who may suffer lasting damage when exposed to harmful content.⁸
- 2. Platforms are only reacting** to reports of harm. This places the burden on the affected users and their willingness and ability to notify platforms of incidents. Yet many users do not submit reports due to shame or fear of repression or feeling overwhelmed. The reluctance to register complaints is often caused by reporting systems that are intentionally designed to confuse users or discourage reports. This is particularly concerning given that many potentially social media-related harms such as anxiety, addiction, or depression cannot even be reported.⁹
- 3. Many platform providers hesitate** to take comprehensive action—even after having been notified about abusive content or behaviour. For example, a recent study by HateAid shows that 55% of the illegal content reported to social media platforms stayed online.¹⁰ Moreover, the inconsistent application of moderation rules confuses users and demonstrates the arbitrary nature of platforms' current safety design.

We must therefore move from reactive damage containment to proactive harm prevention, making safety integral to platform design. This is not a revolutionary approach but common practice in the field of product design safety.



In late 2025, X's AI chatbot Grok added image generation which enabled users to create sexualised images of women and girls, including minors and public figures. Within nine days, Grok generated more than 1.8 million sexualised deepfakes, along with other harmful and extremist content.^{11, 12}

Several countries began investigating or outright banned Grok, prompting X to restrict access behind a paywall.¹³ **Despite X's claims to the contrary, Grok still creates non-consensual sexual deepfakes in the EU, the UK, and the US.**¹⁴ **The company thus continues to facilitate, promote, and capitalise on gender-based violence.**

To change this, HateAid commissioned two expert opinions that provide an in-depth analysis of concrete possible technical and legal measures that minimise the risk of online harms and maximise platform safety. The research is focused on commonly known social media platforms which operate and deploy their functionalities globally.

Furthermore, a comprehensive taxonomy was built that provides a structured overview of hundreds of actionable SbD features and design principles. It includes features that are currently in use worldwide, those that have been used in the past, and those that are not in use but would be recommended. The taxonomy and scientific expertise presented in the following are intended as guidance for regulatory authorities and (emerging) social media platforms on possible technical measures and pre-emptive structural solutions that place user safety at the heart of platforms' design.



Prof. Dr. Michael Denga is a civil and commercial lawyer with a particular focus on the regulation of digital technologies. His research connects core areas of German civil and commercial law with European perspectives, addressing issues related to platforms, data, and artificial intelligence. Since 2025, he has held the Chair of Civil Law and Commercial Law at BSP Business and Law School Berlin.

[The expert report can be found here.](#)



Caroline Sindere is a machine-learning-design researcher and artist. For the past few years, they have been examining the intersections of technology's impact in society, interface design, artificial intelligence, abuse, and politics in digital conversational spaces. Sindere is the founder and director of Convocation Design + Research (CoRD Labs), a human rights research and technology lab.

[The expert report can be found here.](#)

The Principles of Safety by Design

The SbD concept has been researched for more than ten years. While there is no commonly accepted definition, there seems to be a consensus on the underlying principles.¹⁵ For the purposes of this publication, we will rely on the definition proposed by CoRD Labs, a research institute:

“Safety by Design is an approach and methodology, including a related taxonomy, that centers safety for individual users, and also groups, collectives, and communities from the beginning of the technology, design and ideation process of software and hardware development, including and especially focused on very large online platforms.” (CoRD Labs, 2026)

“[S]afety is not the absence of risk, but the presence of conditions that allow people to express themselves, connect with others, and participate in digital spaces with agency.” (CoRD Labs, 2026)

These principles align closely with regulatory frameworks in Australia and the UK, where SbD is embedded in national platform safety regulations.^{17,18} Although the EU's Digital Services Act (DSA) does not explicitly reference SbD, it incorporates similar requirements, such as mitigating systemic risks through platform design,¹⁹ user-friendly reporting channels, and the labelling of automated takedown decisions.²⁰ As SbD principles gain regulatory traction, the focus must now turn to establishing best practices and enhancing the practical implementation of safety obligations.

With regards to the underlying principles, CoRD Labs states that SbD goes beyond isolated security and privacy measures to address all sorts of harms (**Safety**). This requires the consistent evaluation and systemic countering of emerging risks throughout the entire life cycle of a platform (**Proactivity**). To guarantee that certain risks are not overlooked, SbD centres the experiences of those most vulnerable to harm (**Intersectionality**). It also ensures that users understand (**Legibility**) how their data, content, and platform interactions are processed (**Transparency**)—even if they face language or disability barriers (**Accessibility**). Consequently, SbD enables all users to make meaningful choices about their safety, privacy, and participation online (**Agency**).¹⁶

Our Safety-by-Design Taxonomy

On behalf of HateAid, CoRD Labs has created a comprehensive SbD taxonomy. It includes no less than 214 hands-on recommendations (**Interventions**) that are clustered under 30 core SbD principles (**Attributes**), 62 design models (**Design Patterns**), and 39 technical tools (**Components**). In the following, eight illustrative examples of effective SbD measures will be presented.



[The entire taxonomy can be found here.](#)

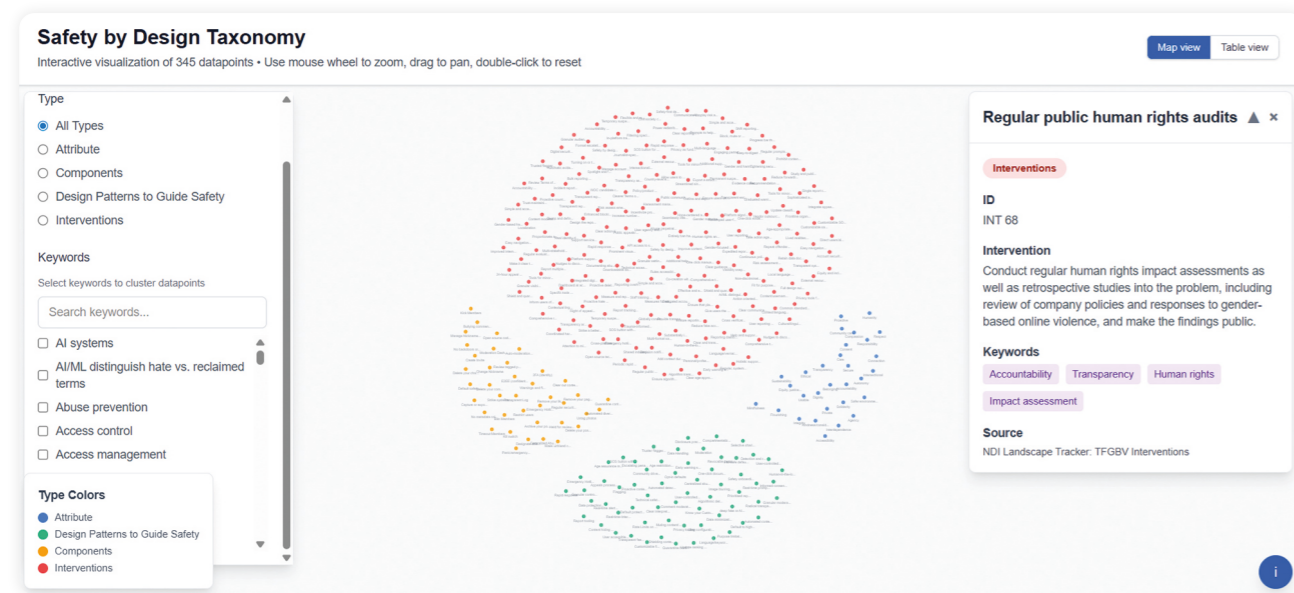


Figure 1 - Screenshot of the Taxonomy. The interactive visualisation consists of 345 data points.

Surprisingly for some, many of the listed interventions come with a proven track record as they have been tried and tested on social media platforms for years. For example:

Friction and User Nudging

Instagram uses automated systems to detect users attempting to post potentially offensive content. It then displays warnings within the app, notifying those users that repeated violations of community guidelines may result in account suspension. Similarly, during the 2020 US election, Twitter introduced a prompt that nudged users to read articles before sharing them to minimise the risk of spreading disinformation. By leveraging such tools, platforms could adopt comparable strategies to proactively deter the spread of harassing language and disinformation from the outset.²¹

Abuse Managing Tools

In 2022, Google's Jigsaw incubator launched an open-source harassment manager for Twitter. The tool was explicitly designed to help women journalists, who often face a high volume of gender-based online violence. Its purpose was to "easily identify and document harmful posts, mute or block perpetrators of harassment and hide harassing replies to their own tweets".²² Due to changes in Twitter/X's operation, the tool is unfortunately no longer functional.²³ A new version of this tool would be an essential asset for those most likely to experience large-scale online abuse.

Safety Modes

Features like X's "Protect my Posts" and Instagram's "Private Mode" enable users to limit the visibility of their profile with just one click. This provides users with greater control over their privacy and protects them from unwanted interactions with strangers.²⁴ It can also function as a panic button for users facing a sudden wave of online abuse. To further enhance usefulness of this feature, platforms should offer customisable safety modes that transcend the simple public/private binary.²⁵

Account Labelling

Social media platforms have long experimented with labels to combat disinformation and improve transparency. Twitter introduced blue check marks back in 2009 to verify public figures and institutions. In 2023, the company, now called X, began selling these check marks, rendering its well-established verification system ineffective. Instead, X added a labelling requirement for parody accounts to identify satirical content.²⁶ YouTube has labelled state-funded news outlets since 2018,²⁷ though inconsistently.²⁸ Platform markers for verified or trustworthy sources can help users assess content credibility and reduce disinformation spread.

Other recommendations include novel approaches that address existing shortcomings in social media platforms' design. For example:

Cross-Platform Cooperation

"Online violence often jumps across platforms and exploits the weaknesses of each."²⁹ Cyberbullies and stalkers, for example, are likely to pursue their victims on multiple platforms simultaneously. Consequently, targeted users would greatly benefit from the ability to block abusive keywords, content, and users across platforms. Cross-platform doxxing alert systems would help to quickly contain the spread of leaked personal information once it has been detected on one platform.

Rapid Response Teams

Social media platforms should let users build rapid response teams comprised of trusted allies who can help tackle online abuse. Users should be able to assign these allies limited account access—akin to Gmail's delegate feature—to monitor, document, and report harassment. To this end, platforms could build on existing features like X's "teams" function (currently limited to X Pro) or Instagram's "roles" (only available for shared Instagram business accounts).³⁰

One-Click Documentation Tool

Online abuse needs to be properly recorded to be prosecuted. Yet users often do not know how to create legally compliant screenshots, and face time pressure as abusive content may be swiftly taken down or deleted. To ensure that online abuse does not go undocumented, platforms could integrate a one-click documentation feature that automatically captures abusive posts, along with essential metadata and jurisdiction-specific details, whenever users report or block content.³¹

Quarantining Content under Review

Information spreads fast on the internet. This includes disinformation, personal data such as a private address, or links to leaked material. Quarantining reported content until it has been reviewed thus constitutes an essential step in decreasing the visibility and containing the spread of harmful information. Victims of large-scale online harassment would also benefit from a quarantine dashboard that allows them to safely scan potentially abusive content with the help of trusted allies.³²

The taxonomy seeks to offer policymakers and platform providers a clear, actionable overview of both established and innovative SbD features, serving as a foundation for building safer social media platforms. However, it is important to keep in mind that SbD is not just about adding features or enforcing compliance. Rather, it requires the proactive and continuous adaptation of safety features to an ever-changing online environment, thereby centring the experience and needs of the users most affected.

"True Safety by Design means anticipating how systems can be weaponized [...]. [It also includes] building protections before the first user ever signs up, and responding to harms in real time to more easily address and mitigate those harms at scale." (CoRD Labs, 2026)

How to Implement Safety-by-Design Measures

Barriers to Implementation

Despite the existence of an extensive catalogue of well-known safety features, why do most social media platforms choose to put their users at risk in-

stead? Based on the input provided by the consulted experts, three key reasons for platforms' unwillingness to adopt SbD measures can be stressed:

Platforms' Business Models

Social media platforms generate revenue by collecting and monetising their users' data. To maximise their profits, they constantly seek to increase the numbers of users and their engagement with the platform. The more a user interacts with a platform—through (dis)likes, comments, and posts—the more data can be collected, and the better ads can be personalised. Therefore, social media platforms promote viral content that is often controversial, untrue, or downright harmful, but highly engaging.³³ They also make use of addictive design choices that keep users glued to their screens.

SbD directly challenges this profit model by putting users' well-being and safety over engagement and screen time. In addition to limiting the reach of polarising and harmful content, SbD also champions privacy and transparency, while empowering users to limit the tracking of their online behaviour. It is not only the financial cost of implementing new safety features, but the prospective loss of user data that conflicts with the commercial interests of large social media platforms.

Power Asymmetries and Bias

Another critical barrier is power asymmetries within platform design processes. Design teams often base decisions on their assumptions about an "average user". This reinforces existing biases and typically neglects the marginalised, less monetisable communities that are at the highest risk of harm.³⁴ Unlike the majority of users, the engineers behind social media platforms are a relatively homogenous group of mostly male tech-savvy and well-paid professionals that share certain values and interests. Contrary to doctors and lawyers, there is neither an oath, nor proof of qualification, nor even an industry code that obliges social media platform engineers. Accordingly, those with the power to design the platforms that shape our public discourse lack clear incentives to take the needs of groups other than their own into account.

Take TikTok's recent decision to replace its trust and safety team with flawed automated moderation systems, for example.³⁵ Though this design change severely exacerbates the risk of harm for users, particularly those belonging to marginalised communities, there was no analysis of the possible consequences for different target groups. By centering design choices on the experiences of those most affected, SbD threatens to subvert this power dynamic.

Lack of Regulatory Incentives

In the EU, most types of consumer products are subject to strict safety standards to prevent harm. If producers fail to comply, they can be held accountable by those damaged by their negligence. By contrast, social media platforms have long enjoyed broad liability privileges to ensure the free flow and amplification of user-generated content, while facing comparatively few safety requirements to attract service providers to the EU. After two decades of minimal regulation, the EU adopted the DSA, which sets stricter safety standards. The landmark law also specifies that social media platforms are only exempt from liability for user-generated content until it has been reported.³⁶ But whereas the DSA introduces single SdD measures, it lacks a comprehensive underlying SbD mandate that binds individual measures together.

This has allowed platforms to undermine obligations through superficial and partial compliance. For example, users are continuously discouraged to report harmful content because of the design of reporting channels,³⁷ while a significant part of reported content stays online anyway.³⁸ Neither platforms' transparency reports nor the European Commission's transparency database provide much insight into the inner workings of recommender systems or the quality of content moderation. Vetted researchers requesting data access to assess platforms' processes frequently receive information that is too generic or unstructured for scientific purposes.³⁹ External audits assessing systemic risks, such as public health or election security, remain superficial, incomplete, and lack valid evidence.⁴⁰ Likewise, most large platform providers have shown little interest in properly assessing and mitigating the systemic risks stemming from their services.

Though some European and national enforcement authorities have started to investigate and fine social media platforms for their poor compliance, significant enforcement deficits remain. On the one hand, regulatory bodies simply lack the staff, expertise, and funding to process and prosecute the huge volume of reports of platform maladministration in a timely manner. This problem is exacerbated by overlapping mandates and regulatory frameworks that necessitate lengthy coordination processes between different jurisdictions and enforcement bodies in charge of platform regulation, data protection, artificial intelligence, copyright protection, and media oversight. On the other hand, enforcement authorities face increasing political backlash, as platform providers and the US administration seek to pressure the EU into repealing its digital laws. Thus, enforcement of the DSA has now become a bargaining chip in geopolitical conflicts, with compliance leveraged as a tool in broader transatlantic disputes. The future of Europe's digital sovereignty hinges consequently on the willingness and ability of regulators to enforce their rulebook, having to stand up against US-backed platforms which are unafraid of violating the legislation of the countries they are doing business in.⁴¹

Recommendations for Policymakers

Europe's approach to platform regulation is uniquely rooted in its commitment to democratic values, the rule of law, and fundamental human rights. To defend these principles and safeguard its digital sovereignty, the EU must assert the user rights enshrined in its legislation—resisting both the financial interests of platform providers and transatlantic political pressure. In his legal expert opinion, Prof. Denga emphasises that this is not merely a political choice but a legal obligation, comparable to the EU Commission's duty to enforce competition law. He argues that legislators must fully implement the DSA, as “pursuing effective European digital sovereignty requires a significant reduction in the

discretion of the commissioned authorities, which may oblige to intervene into business models.”⁴²

While the DSA provides a robust legal foundation, its success depends on moving beyond compliance alone. A holistic SbD approach is essential to address systemic risks that the DSA alone cannot fully mitigate. To support regulators in this transition, we have developed a three-step programme that operationalises SbD, ensuring that Europe's digital infrastructure aligns with its core values and sets a global standard for responsible social media platform governance.

Step 1: Resolute Enforcement of Existing Rules

The withdrawal of the TikTok Lite Rewards programme, which encouraged addictive behaviour through monetary vouchers and coins, demonstrates how strict and swift enforcement can prevent large-scale harm.⁴³ Consequently, the most immediate step centres the robust enforcement of the DSA, particularly its design-related elements. This includes ensuring that social media platforms create user-friendly reporting channels and remove illegal content consistently and promptly. It must also ensure that researchers are provided with useful data, insightful transparency reports are compiled, and that social media platforms are vetted by independent auditors. Enforcement authorities thus urgently need more financial and human resources to monitor and ensure compliance. They must also be shielded from political interference. To ensure the latter, an independent enforcement body on the European level is recommended. This may be a new European agency or even an entirely new EU body that would—like the European Central Bank—exercise its oversight without direct political guidance.⁴⁴

The backing of the US government may embolden certain platform providers to openly defy European regulation. In these cases, it is important to remember that “[p]latform obligations are enforced not only through regulatory and private law liability of the providers of intermediary services, but also through the personal liability of their staff.”⁴⁵ Although DSA obligations are directed at platforms, not staff, Prof. Denga points out that a personal responsibility for compliance arises. It is based on the objective quality standards of intermediary services set by the DSA, the pressing public interest enshrined in the principle of European digital sovereignty, and the benefit-burden paradigm of liability law.⁴⁶ Consequently, CEOs and decision-makers at social media platforms could be held personally liable in the event of serious damage caused by systemic failures, particularly if gross negligence or intentional noncompliance can be proven.⁴⁷

Step 2: Introduction of Actionable Safety Standards

The European Commission,⁴⁸ the European Parliament,⁴⁹ and the European Council⁵⁰ all have argued that the current safety standards are insufficient to properly protect adult consumers and children online. Accordingly, new and updated legislation is required to outlaw, among others, the creation and dissemination of nonconsensual deepfakes, addictive design features such as autoplay or endless scrolling, and practices like the use of social bots and shadowbanning that distort public discourse.⁵¹ In addition to banning manipulative or harmful practices, policymakers must also proactively mandate specific safety standards and measures. This includes classifying social media algorithms and content-creating AI tools, such as Grok, as high-risk AI systems to ensure their compliance with transparency obligations like data quality checks or bias monitoring.⁵² Other actionable examples can be found among the 214 interventions listed in the SbD taxonomy, which may serve as a point of reference.⁵³

To prevent superficial compliance, new legislation needs to underpin individual safety measures with an overarching SbD mandate. This entails obliging social media company executives and local representatives to fulfil “specific qualification requirements [...] to ensure they are able to meet the challenges of digital opinion spaces, similar to the ‘fit and proper’ requirements in capital market law”.⁵⁴ Moreover, the mandate should introduce procedural safeguards in platforms’ design processes, such as regular stakeholder consultations, to guarantee their safety, transparency, legibility, accessibility, intersectionality, and focus on user agency. Given the recent surge of SLAPP lawsuits by social media platforms, online violence, and even acts of repression by foreign governments, stronger protections for individual researchers and civil society organisations against such attacks are also called for.

Step 3: Decentralisation of the Social Media Infrastructure

For the past 20 years, social media companies have designed their platforms to be as attention-grabbing and engaging as possible. This allowed a few small tech startups to turn into powerful multinational corporations. It also led to the proliferation of digital violence and online disinformation. Instead of taking responsibility, platform providers deflect blame and shift the burden on law enforcement agencies with whom they cooperate only to a limited extent. While they would be perfectly capable of adapting their platforms’ design to ensure greater safety, most social media companies are unwilling to hurt their profits.

Consequently, ever more people withdraw from public discourse and social divisions are being exacerbated for the benefit of a few companies. This has dire implications for our democracy and security. Policymakers should explore ways to promote alternative models of social media. Existing decentralised, open-source networks such as the Fediverse or Eurosky offer alternatives to established platforms and operate without surveillance-based advertising or addictive algorithms.⁵⁵ However, these noncommercial platforms struggle to attract private capital. Therefore, more public funding for European social media initiatives and the introduction of legally enshrined portability rights that facilitate user migration between platforms are needed. Moreover,

the existing liability regime for hosting providers is not designed to deal with decentralised networks and poses legal challenges and insecurities. Therefore, the EU needs to actively consider special rules to make them thrive, which is hardly possible in a framework made to regulate the most powerful and abusive players. Exemptions for noncommercial platforms and more regulatory sandboxes are needed.

By championing the development and widespread adoption of European social media models that embrace the SbD principles mentioned in the taxonomy, the EU can effectively counter systemic online harms while affirming its digital sovereignty. As the European Parliament’s recent topical resolution noted, digital sovereignty is not just a technical challenge but a democratic imperative.⁵⁶ In an era where digital spaces shape political realities, democratic control over social media platforms is essential to uphold European autonomy and fundamental values. The time has come for the EU to turn its principles into action and build a digital future that reflects its commitment to safety, accountability, and the primacy of the rule of law.

List of Sources

- 1 Eurobarometer (2025): Social Media Survey 2025 (FL014EP), p. 12, <https://europa.eu/eurobarometer/surveys/detail/3592> (accessed: 31/01/2026).
- 2 Bhargava, Vikram R. & Manuel Velasquez (2021): Ethics of the Attention Economy: The Problem of Social Media Addiction, *Business Ethics Quarterly*, 31 (2021), 321–59. DOI: 10.1017/beq.2020.32.
- 3 Munn, Luke (2020): Angry by design: toxic communication and technical architectures, *Humanities and Social Sciences Communications* 7(1): 1–11. DOI: 10.1057/s41599-020-00550-7.
- 4 Mujica, Alejandro L., Crowell, Charles R.; Villano Michael A. & Uddin, Khutb M. (2022): Addiction by design: Some dimensions and challenges of excessive social media use, *Medical Research Archives*, 10(2), 1–29. DOI: 10.18103/mra.v10i2.2677.
- 5 Lee, Hae Yeon; Jamieson, Jeremy P.; Reis, Harry T.; Beevers, Christopher G.; Josephs, Robert A.; Mullarkey, Michael C.; O'Brien, Joseph & Yeager, David S. (2020): Getting fewer "Likes" than others on social media elicits emotional distress among victimized adolescents, *Child Development*, 91(6), 2141–2159. DOI: 10.1111/cdev.13422.
- 6 Horwitz, Jeff (2025): Meta buried 'causal' evidence of social media harm, US court filings allege, Reuters, <https://www.reuters.com/sustainability/boards-policy-regulation/meta-buried-causal-evidence-social-media-harm-us-court-filings-allege-2025-11-23/> (accessed: 31/01/2026).
- 7 Conger, Kate; Freedman, Dylan & Thompson, Stuart A. (2026): Musk's Chatbot Flooded X With Millions of Sexualized Images in Days, *New Estimates Show*, *New York Times*, <https://www.nytimes.com/2026/01/22/technology/grok-x-ai-elon-musk-deepfakes.html> (accessed: 31/01/2026).
- 8 McHugh, Bridget Christine; Wisniewski, Pamela; Rosson, Mary Beth & Carroll, John M. (2018): When social media traumatizes teens: The roles of online risk exposure, coping, and post-traumatic stress. *Internet Research*, 28(5), 1169–1188. DOI: 10.1108/IntR-02-2017-0077.
- 9 Das Netz (2025): Between Click and Consequence: An Evaluation of Platform Reporting Procedures under the Digital Services Act, https://www.das-netz.de/sites/default/files/2025-11/ENG_NETTZ_Meldewege_18.pdf (accessed: 16/12/2025).
- 10 HateAid (2025): Rights Without Reach. The DSA Put to the Test, <https://hateaid.org/wp-content/uploads/2025/12/hateaid-dsa-rights-without-reach-2025.pdf> (accessed: 31/01/2026).
- 11 CCDH (2026): Grok floods X with sexualized images of women and children, <https://counterhate.com/research/grok-floods-x-with-sexualized-images/> (accessed: 31/01/2026).
- 12 Bouchau, Paul (2026): Grok Generating Flood of Sexualized Images of Women and Minors, *AI Forensics*, <https://aiforensics.org/work/grok-unleashed> (accessed: 31/01/2026).
- 13 Watkins, Ali (2026): Malaysia and Indonesia Block Access to Grok Because of Sexually Explicit Content, *New York Times*, <https://www.nytimes.com/2026/01/11/world/asia/malaysia-indonesia-grok-ban.html> (accessed: 31/01/2026).
- 14 Lyons, Emmet (2026): X, Grok AI still allow users to digitally undress people without consent, as EU announces investigation, *CBS News*, <https://www.cbsnews.com/news/x-grok-ai-imagery-elon-musk-eu-uk-us-regulation/> (accessed: 31/01/2026).
- 15 Woods, Lorna (2024): Safety by Design, <https://www.onlinesafetyact.net/analysis/safety-by-design/> (accessed: 31/01/2026).
- 16 Sindera, Caroline; Valencia, Antonia & Smith, Sam (2026): Safety by Design: A comprehensive methodology and exploration on design, policy and technology interventions to generate better user safety on platforms, p. 24–29, <https://hateaid.org/wp-content/uploads/2026/03/safety-by-design-research-report-platform-safety-interventions-caroline-sindera.pdf> (accessed: 31/01/2026).
- 17 eSafety Commissioner (2025): Safety by Design, <https://www.esafety.gov.au/industry/safety-by-design> (accessed: 31/01/2026).
- 18 The Secretary of State for Science, Innovation and Technology (2025): Final Statement of Strategic Priorities for Online Safety, www.gov.uk/government/publications/statement-of-strategic-priorities-for-online-safety/final-statement-of-strategic-priorities-for-online-safety#-safety-by-design (accessed: 31/01/2026).
- 19 Cf. Art. 35(1)(a), DSA.
- 20 Cf. Art. 16 ff, DSA.
- 21 Jankowicz, Nina; Hunchak, Jillian; Pavliuc, Alexandra; Davies, Celia; Pierson, Shannon & Kaufmann, Zoë (2021): Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women Online, *Wilson Center*, <https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online> (accessed: 31/01/2026).
- 22 Jigsaw (2022): Technology to help women journalists document and manage online abuse, *Medium*, <https://medium.com/jigsaw/technology-to-help-women-journalists-document-and-manage-online-abuse-5edcac127872> (accessed: 31/01/2026).
- 23 Thomas Reuter Foundation (2025): TRFilter, <https://www.trfilter.org/> (accessed: 31/01/2026).
- 24 Instagram automatically enables its private mode for all underaged users to prevent grooming.
- 25 Chumsky, Susan (ed.) (2021): No Excuse for Abuse. What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users, *PEN America*, <https://pen.org/report/no-excuse-for-abuse/> (accessed: 31/01/2026).
- 26 Maxwell, Thomas (2025): X Creates a New Parody Label for Accounts, Solving a Problem Elon Created, *Gizmodo*, <https://gizmodo.com/x-creates-a-new-parody-label-for-accounts-solving-a-problem-elon-created-2000548621> (accessed: 31/01/2026).
- 27 BBC (2018): YouTube to label government and public-funded clips, <https://www.bbc.com/news/technology-46139189> (accessed: 31/01/2026).
- 28 Kofman, Ava (2019): YouTube Promised to Label State-Sponsored Videos But Doesn't Always Do So, *ProPublica*, <https://www.propublica.org/article/youtube-promised-to-label-state-sponsored-videos-but-doesnt-always-do-so> (accessed: 31/01/2026).
- 29 Sindera et al. (2026), p. 61.
- 30 Chumsky (2021).
- 31 Ibid.
- 32 Ibid.
- 33 Hagey, Keach & Horwitz, Jeff (2021): Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead., *The Wall Street Journal*, <https://www.wsj.com/tech/facebook-algorithm-change-zuckerberg-11631654215> (accessed: 31/01/2026).
- 34 Costanza-Chock, Sasha (ed.) (2020): *Design Justice: Community-Led Practices to Build the Worlds We Need*, *The MIT Press*. DOI: 10.7551/mitpress/12255.001.0001.
- 35 Kerr, Dara (2025): TikTok to replace trust and safety team in Germany with AI and outsourced labor, *The Guardian*, <https://www.theguardian.com/technology/2025/aug/10/tiktok-trust-safety-team-moderators-ai> (accessed: 31/01/2026).
- 36 European Union (2022): Regulation (EU) 2022/2065 (Digital Services Act), Art. 4–6, O.J. (L 277) 1.
- 37 Bösward, Lena-Maria; Dolezalek, Corinna; Jost, Pablo & Schmid, Ursula Kristin (2025): Between Click and Consequence: An Evaluation of Platform Reporting Procedures under the Digital Services Act, *Das Netz*, https://www.das-netz.de/sites/default/files/2025-10/ENG_Langfassung_DSA.pdf (accessed: 31/01/2026).
- 38 HateAid (2025): Rights Without Reach. The DSA Put to the Test, <https://hateaid.org/wp-content/uploads/2025/12/hateaid-dsa-rights-without-reach-2025.pdf> (accessed: 31/01/2026).
- 39 Denga, Michael (2025): Enforcement and optimization of social media platforms' responsibility, p. 18–19, <https://hateaid.org/wp-content/uploads/2026/03/safety-by-design-expert-opinion-platform-responsibility-michael-denga.pdf> (accessed: 31/01/2026).
- 40 Holznagel, Daniel (2025): Shortcomings of the first DSA Audits – and how to do better, *DSA Observatory*, <https://dsa-observatory.eu/2025/06/11/shortcomings-of-the-first-dsa-audits-and-how-to-do-better/> (accessed: 31/01/2026).
- 41 Denga (2025), p. 10–12.
- 42 Ibid., p. 12.
- 43 European Commission (2024): TikTok commits to permanently withdraw TikTok Lite Rewards programme from the EU to comply with the Digital Services Act, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4161 (accessed: 31/01/2026).
- 44 Harfst, Jan-Ole, Mast, Tobias & Schulz, Wolfgang (2025): Independence as a Desideratum: DSA Enforcement by the EU Commission, *Verfassungsblog*, <https://verfassungsblog.de/dsa-enforcement-commission/>, DOI: 10.59704/50020240f8894397.
- 45 Denga (2025), p. 14.
- 46 Ibid., p. 6–7, 15–17.
- 47 Ibid., p. 17.
- 48 European Commission (2024): Commission Staff Working Document Fitness Check on EU consumer law on digital fairness (SWD(2024) 230 final), https://commission.europa.eu/document/download/707d7404-78e5-4aef-acfa-82b4cf639f55_en (accessed: 31/01/2026).
- 49 European Parliament (2023): European Parliament resolution of 12 December 2023 on addictive design of online services and consumer protection in the EU single market (2023/2043(INI)), https://www.europarl.europa.eu/doceo/document/TA-9-2023-0459_EN.html (accessed: 31/01/2026).
- 50 Council of the European Union (2025): The Jutland Declaration: Shaping a Safe Online World for Minors, https://www.digmin.dk/Media/638956829775203140/DIGMIN_The%20Jutland%20Declaration%20Shaping%20a%20Safe%20Online%20World%20for%20Minors%20101025.pdf (accessed: 31/01/2026).
- 51 Denga (2025), p. 20–22.
- 52 Ibid., p. 9–10.
- 53 Convocation Research + Design (2026): Safety by Design Taxonomy, <https://sbd-taxonomy.vercel.app/> (accessed: 31/01/2026).
- 54 Denga (2025), p. 21.
- 55 Penfrat, Jan (2022): Everyone is on Mastodon now, but why?, *EDRI*, <https://edri.org/our-work/everyone-is-on-mastodon-now-but-why/> (accessed 31/01/2026).
- 56 European Parliament (2026): European Parliament resolution of 22 January 2026 on European technological sovereignty and digital infrastructure (2025/2007(INI)), https://www.europarl.europa.eu/doceo/document/TA-10-2026-0022_EN.html (accessed: 31/01/2026).

Imprint

HateAid gGmbH
Greifswalder Straße 4
10405 Berlin

Email: kontakt@hateaid.org
hateaid.org

Company headquarters: Berlin
Register court: Charlottenburg Local Court
Commercial register number: HRB 203883 B
VAT ID number: DE322705305

EU Transparency Register ID: 802412042190-08

Managing Directors and Responsible for Editorial Content: Anna-Lena von Hodenberg, Josephine Ballon

